Visit our website

A powerful multi-omic AI framework for blood-based cancer screening and tissue of origin prediction

Badri Padhukasahasram*, Lavanya Nemani*, Sayak Dhar*, Shekhar Nath, Shardul Kamble, Advika Lakshman, Keshav Agarwal, Rohan Karthikeyan, Krithika Subramanian, Sarath Ra, Kannadhasan Sekar, Shashank Kumar, Prakash Hiremath, Nihesh Rathod, Neelkanth Mishra, Stan Iyer, Swaraj Basu, Radhakrishna Bettadapura, Vamsi Veeramachaneni, Ramesh Hariharan

Strand Life Sciences, Bangalore, India *Contributed equally

Contact

Badri Padhukasahasram
Vice President - Data Science

+91-9591670961 badri.p@strandls.com

Introduction

- Inferring signatures of cancer in blood is an inherently difficult problem because circulating tumor-DNA is at very low abundance especially during early stages.
- Many current approaches analyze only one modality at a time and depend on task-specific modeling strategies.
- We present a unified multi-omic AI framework that integrates methylation and fragmentomic signals for both cancer detection and tissue-of-origin (TOO) prediction using a tabular data foundation model.
- Our method also incorporates a methylation foundation model embeddings that captures co-methylation structure from a large and diverse corpus.
- Performance demonstrated on an Indian cohort and compared with fragmentomics only and methylation only base lines.

Datasets

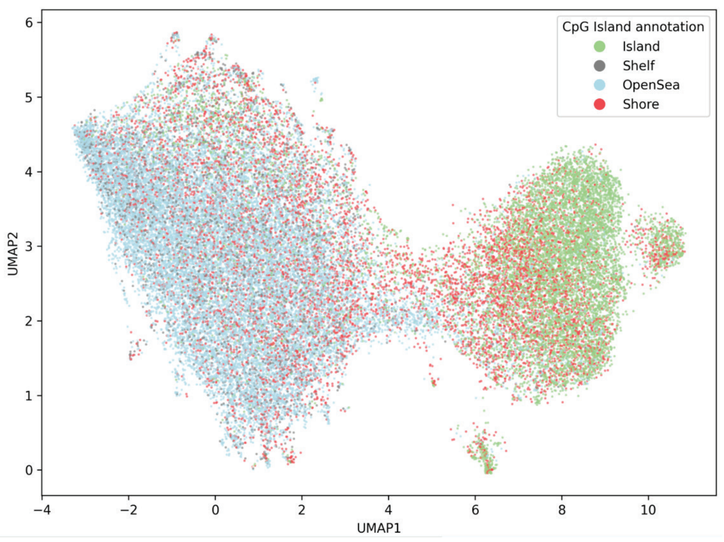
All analyses described here were based on these 2 datasets:
We used 1,014 deidentified cfDNA samples from an Indian cohort (688 cancer cases and 326 controls), sequenced by EM-seq on the Twist methylome panel (~123Mb and ~4 million CpG sites). Each site includes genomic coordinates (hg38) and fragment-level methylation counts (methylated, unmethylated, undetermined). A minimum coverage threshold of 30 reads per CpG site was applied to ensure data quality.

For representation learning, we pretrain a transformer-based methylation foundational model on public tissue methylation dataset comprising of ~75,000 samples from more than 1,500 studies across diverse tissues and conditions. The dataset is across multiple Illumina platforms (27k, 450k, EPIC, EPIC+, EPICv2), enabling robust learning of co-methylation patterns.

Pre-training methylation foundational model

10k CpG-site contexts are represented using (i) β-value (methylation fraction) embeddings, (ii) DNA sequence embeddings from Nucleotide Transformer v2 (500M, multi-species) using ±1 kb sequence context, and (iii) positional embeddings comprising sinusoidal genomic position and chromosome-specific RoPE. Performer model consists of 6 transformer layers (embedding dimension = 256, 8 attention heads) with ~5.2M parameters.

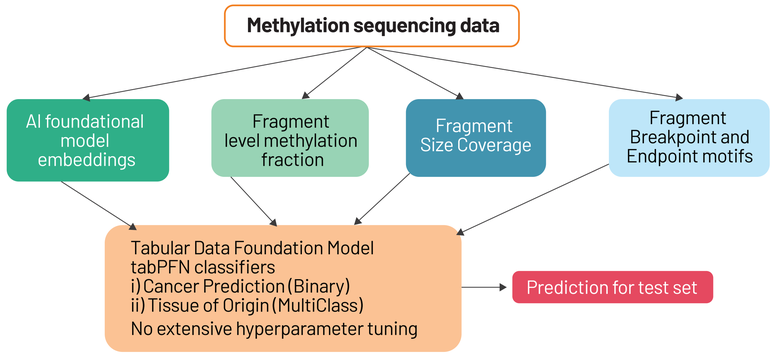
Pre-training: ~15% of CpG sites are randomly masked, and the model reconstructs their β-values from the surrounding context. The learned embeddings capture genome-wide co-methylation and regulatory context, with qualitative analyses suggesting implicit encoding of biologically meaningful features such as CpG islands.



Methods

We extract diverse types of cfDNA-derived features capturing methylation and fragmentation biology:

- Methylation foundation model embeddings sample-level
- Fragment size coverage ratios across genomic windows reflecting tumor-associated shortening.
- End-motif and breakpoint-motif frequencies reflecting altered nuclease activity.
- Fragment based methylation feature aggregated at a region level.



Cancer detection performance by stage

Method	I	II	III	IV
DELFI	65.52%	64.86%	72.97%	71.05%
Fragment level methylation	82.76%	86.49%	86.49%	89.47%
Fragmentation motifs	82.76%	83.78%	78.38%	84.21%
Multi-omic tabPFN	89.7%	91.7%	83.8%	94.7%

- For cancer classification, multi-omic foundation model based AI approach outperformed single modality baselines and provided highest sensitivity at 99% specificity of ~90%.
- Sensitivity tends to improve with cancer stage and stage 4 has highest sensitivity.

Tissue of Origin detection performance

- TOO2 measures the fraction of times true tissue of origin was in the top 2 predicted tissues of origin with highest probabilities from classifier. Our method achieved an accuracy of ~75.7% across 10 different cancer types, and a TOO2 of ~87.9%.
- Both accuracy and TOO2 for multi-omic method higher than alternate methods based only on methylation or fragmentomics.
- Tissues sharing developmental origins may retain similar epigenetic landscapes and could be harder to separate. E.g. foregut-derived tissues such as stomach vs esophagus, pancreas vs gallbladder and lung vs esophagus etc

Method	Accuracy	TOO2
DELFI	51.55%	68.04%
Fragment methylation	69.63%	77.87%
Fragmentation motifs	69.37%	83.78%
Multi-omic tabPFN	75.70%	87.90%

Summary and Achievements

- Integrated multi-omic AI framework achieved strong performance for an indian cohort in terms of both sensitivity and specificity for early cancer detection. F1 score improved by ~15-20% compared to single modality predictors.
- Tissue of origin performance also improved by ~10-30% with use of multi-omic AI model compared to single modality baselines.
- Pre-trained methylation foundation model generates biologically meaningful embeddings that can be further improved by training for longer number of epochs.

Conclusions

- Learned methylation representations using a CpG masking and imputation task. Despite its relatively small size (~5M parameters), the model achieved competitive performance (Pearson coefficient of 0.8919 vs 0.9952 for larger models with ~100M parameters), indicating efficient learning of co-methylation structure.
- CpG sites cluster by genomic context (islands, shores, open sea), demonstrating biological relevance.
- Integrating foundation model-based methylation representations with complementary fragmentomic features provides a powerful and scalable approach for cfDNA-based cancer screening and tissue of origin detection.
- Combining tissue-specific regulatory signals with genome-wide fragmentation patterns significantly improved early cancer detection and tissue of origin significantly as compared to single modality baselines.

Computational Resources

- Our methylation foundational model (~5.2M parameters; 6 transformer layers, embedding dimension = 256, 8 attention heads) has been pre-trained for 8 epochs (~4.5 hrs / epoch) on an NVIDIA A10G GPU; cost ~200$.
- Fine-tuning with our multi-omic modelling approach for cancer detection and tissue of origin detection on RTX 3060 GPU takes ~30 mins and ~40 mins respectively. cost ~0.6$.

References

- DELFI:Genome-wide cell-free DNA fragmentation in patients with cancer. Nature 2019
- Performer: Choromanski, Krzysztof, et al. "Rethinking Attention with Performers." arXiv preprint arXiv:2009.14794 (2020).
- RoPE: Su, Jianlin, et al. "RoFormer: Enhanced Transformer with Rotary Position Embedding." Neurocomputing 568 (2024): 127063.
- Nucleotide-transformer-v2-500m-multi-species: Dalla-Torre, H., et al. "Nucleotide Transformer: Building and Evaluating Robust Foundation Models for Human Genomics." Nature Methods 22 (2025): 287–297.
- Fragment size coverage:Early detection of multiple cancer types using multidimensional cell-free fragmentomics. Nature Medicine 2025